Precise, Fast, and Low-cost Concept Erasure in Value Space: Orthogonal Complement Matters

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Wang Xiang, Xiangnan He

Presented by: Yuan Wang

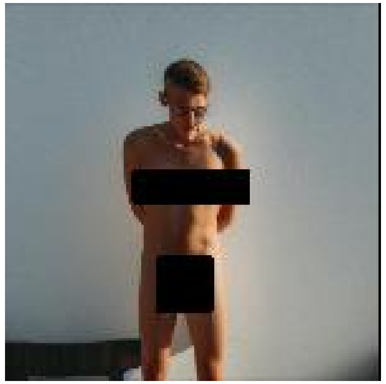
1. Motivation

Practical Needs of Concept Erasure

Parodying IP characters

Infringement of art styles

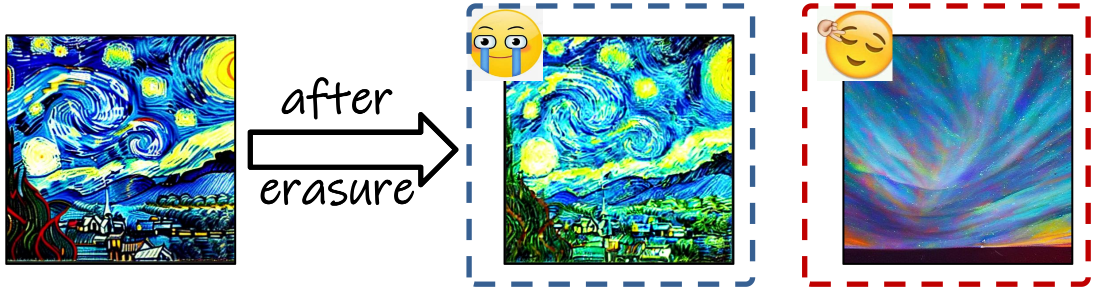
Mocking political celebrities



Unsafe content generation

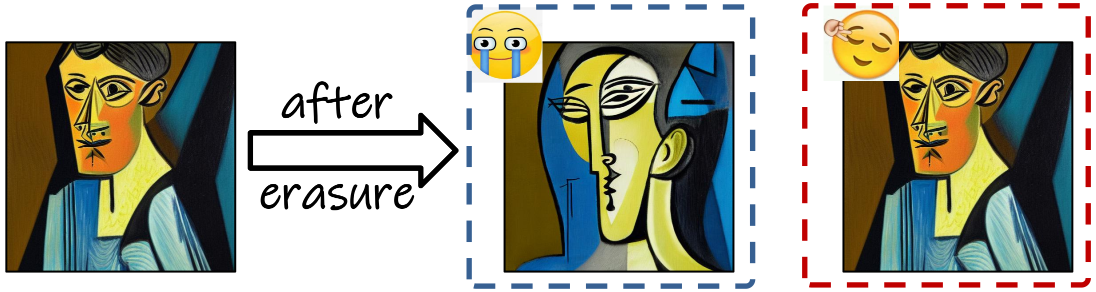
Twofold Demand of Concept Erasure

Target concept: Erasure Efficacy



Precisely erase visual content aligned with target concepts during generation

Non-target concept: Prior Preservation



Minimal impact on non-target content generation.

1. Motivation

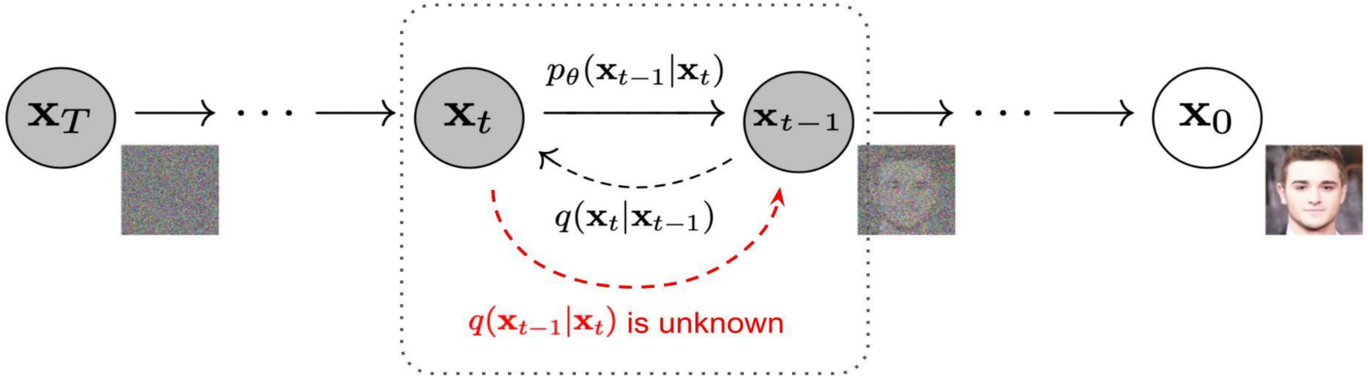
■ Drawbacks of Current Erasure Method

- **Re-training & Blocking**: exclude certain training data and retrain the model or block the prompts of concerns and restrict the outputs of concerns ***Safety Checker, Prompt List, ...***
  - × time cost, requires specialized detectors, introduce biases
  - × fragile and easy to bypass
- **Training-based**: finetune a pre-trained generative model, teaching it to "forget" a target concept ***ConAbl, ESD, SPM, MACE, ...***
  - × hard to achieve real-time erasure
  - × fall short in balancing precise erasure and prior preservation.
- **Training-free**: Intervene in the generation process without requiring additional training ***Negative prompt, SLD, SuppressEOT***
  - × lacks fine-grained control over target concepts and compromises prior preservation
  - × fail to achieve system-wide erasing tasks requiring full automation

We need a precise, fast, and low-cost concept erasure method which can achieve not only precise concept erasure but also satisfactory prior preservation.

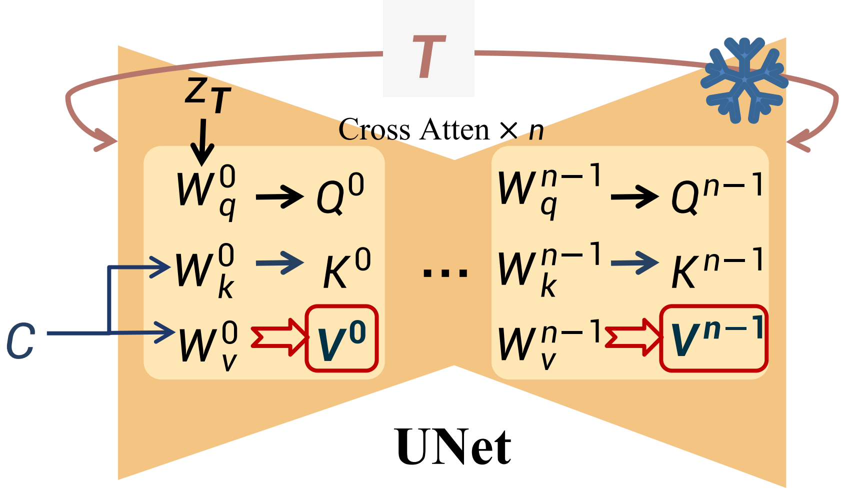
2. Preliminary

- **Stable Diffusion: VAE + UNet**



Sequential denoising of a randomly sampled noisy latent variable is guided by a text prompt using a UNet in the latent space, which can be denoted as $\varepsilon_\theta(z_t, t, C)$.

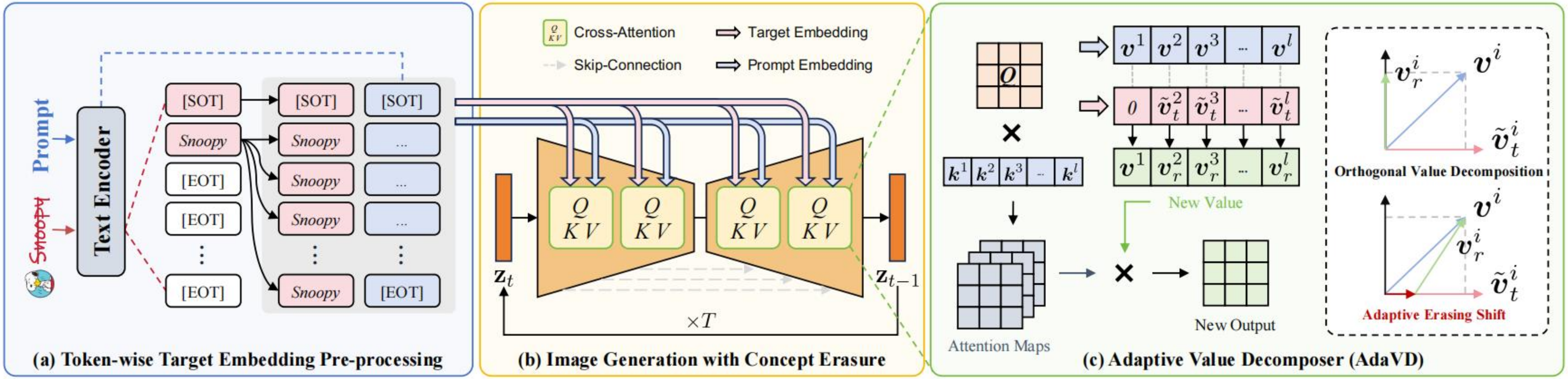
- **Cross Attention Within UNet: align the image with text prompt**



- **Keys** act as the **"Where"** pathway, shaping layout of attention map and image structure.
- **Values** form the **"What"** pathway, controlling the content and appearance of the generated image.

3. Method

■ **Overview of Adaptive Value Decomposer (AdaVD)** **Precise, fast and low cost**

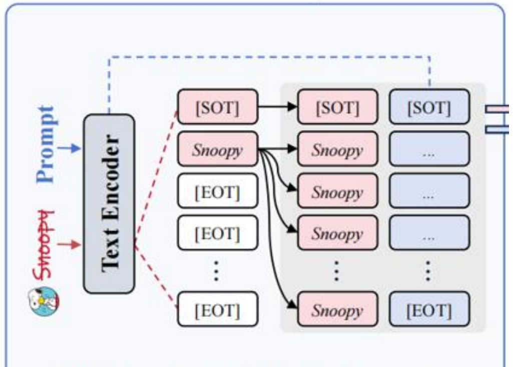


(a) Token-wise Target Embedding Pre-processing (b) Image Generation with Concept Erasure (c) Adaptive Value Decomposer (AdaVD)

- **Step 1**: Token-wisely pre-processing the target embedding of each concept to ensure precise elimination of token-specific target semantics.
- **Step 2**: Feed the pre-processed target embedding and corresponding prompt embedding into CA layers to disentangle target semantics from the original image at each timestep.
- **Step 3**: Perform token-wise orthogonal value decomposition with an adaptive token-specific shift. The new value is subsequently multiplied by the attention map, producing the erased output for this CA layer.

3. Method

Target Embedding Pre-processing



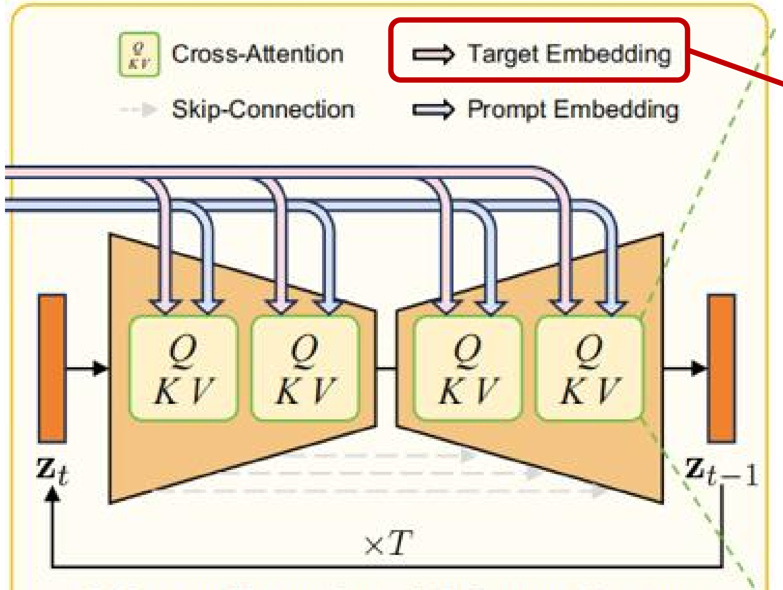
(a) Token-wise Target Embedding Pre-processing

The embedding of the last subject token within its prompt content, which contains tokens excluding [SOT] and [EOT], is duplicated for all the token positions except for [SOT].

E.g. 1: the single-token concept "snoopy" to "[SOT], snoopy, snoopy, ..., snoopy"

E.g. 2: the multi-token concept "Van Gogh" to "[SOT], gogh, gogh, ..., gogh".

Orthogonal Value Decomposition



(b) Image Generation with Concept Erasure

$$\tilde{\mathbf{V}}_t = \tilde{\mathbf{C}}_t \mathbf{W}_{\mathrm{V}} = \left[c_t^1, \underbrace{c_t^k, \dots, c_t^k}_{l-1} \right]^T \mathbf{W}_{\mathrm{V}}. \quad \Longrightarrow \quad \mathbf{V}_t = \left[\mathbf{0}, \tilde{v}_t^2, \tilde{v}_t^3, \dots, \tilde{v}_t^l \right]^T$$

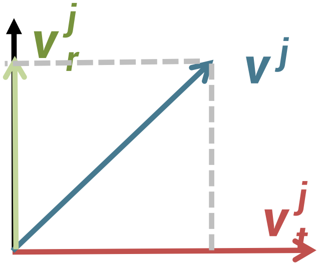
Our proposed erasing operation works by projecting the original text prompt onto the orthogonal complement of the subspace spanned by the target concepts to erase, and it is implemented in the value space learned at each CA layer of the UNet. It supports both single-concept and multi-concept erasure.

3. Method

Orthogonal Value Decomposition

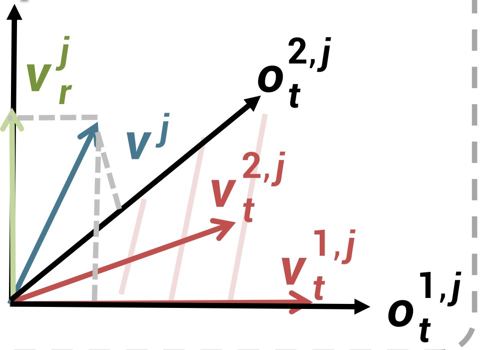
Single-Concept

$$\begin{aligned} \boldsymbol{v}_r^j &= \mathbf{P}_{\mathrm{span}^{\perp}(\boldsymbol{v}_t^j)} \boldsymbol{v}^j = \left(\mathbf{I}_d - \mathbf{P}_{\mathrm{span}(\boldsymbol{v}_t^j)}\right) \boldsymbol{v}^j \\ &= \boldsymbol{v}^j - \frac{\boldsymbol{v}_t^j \boldsymbol{v}_t^{jT}}{\boldsymbol{v}_t^{jT} \boldsymbol{v}_t^j} \boldsymbol{v}^j = \boldsymbol{v}^j - \frac{\boldsymbol{v}_t^{jT} \boldsymbol{v}^j}{\boldsymbol{v}_t^{jT} \boldsymbol{v}_t^j} \boldsymbol{v}_t^j, \end{aligned}$$



Multi-Concept

$$\begin{aligned} \boldsymbol{v}_r^j &= \mathbf{P}_{\mathrm{span}^{\perp}(\{\boldsymbol{v}_t^{h,j}\}_{h=1}^n)} \boldsymbol{v}^j = \mathbf{P}_{\mathrm{span}^{\perp}(\{\boldsymbol{o}_t^{h,j}\}_{h=1}^n)} \boldsymbol{v}^j \\ &= \left(\mathbf{I}_d - \mathbf{P}_{\mathrm{span}(\{\boldsymbol{o}_t^{h,j}\}_{h=1}^n)}\right) \boldsymbol{v}^j \\ &= \boldsymbol{v}^j - \sum_{h=1}^{n} \left(\boldsymbol{o}_t^{h,j}\right)^T \boldsymbol{v}^j \boldsymbol{o}_t^{h,j}. \end{aligned}$$



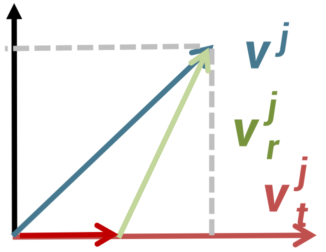
Project the original value vector v^j onto the orthogonal complement of the span of the erased value vector v_t^j or $\{v_t^{h,i}\}_{h=1}^n$ for each token position.

Adaptive Erasing Shift

Single-Concept

$$\boldsymbol{v}_r^j = \boldsymbol{v}^j - \frac{\delta\left(\boldsymbol{v}_t^j, \boldsymbol{v}^j\right) \boldsymbol{v}_t^{jT} \boldsymbol{v}^j}{\boldsymbol{v}_t^{jT} \boldsymbol{v}_t^j} \boldsymbol{v}_t^j.$$

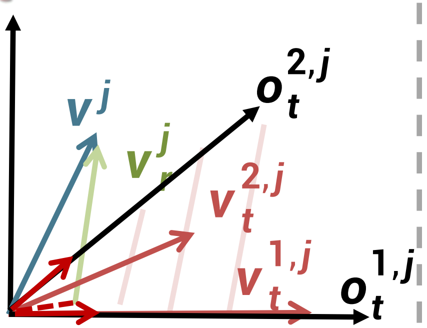
For single-concept, it is straightforward to introduce the shift factor.



Multi-Concept

$$\boldsymbol{v}_r^j = \boldsymbol{v}^j - \sum_{h=1}^{n} \delta\left(\boldsymbol{v}_t^{h,j}, \boldsymbol{v}^j\right) \left(\sum_{k=1}^{n} w_{hk} \left(\boldsymbol{o}_t^{k,j}\right)^T \boldsymbol{v}^j\right) \boldsymbol{v}_t^{h,j}$$

For multi-concept, the orthonormal basis doesn't carry meaningful info.



$$\delta(\boldsymbol{x}, \boldsymbol{y}) = \frac{s}{1 + e^{-p(\cos(\boldsymbol{x}, \boldsymbol{y}) - \epsilon)}}.$$

- ε: quantify and filter the relatively weak relevance
- s: control the factor scale
- p: control the increasing rate

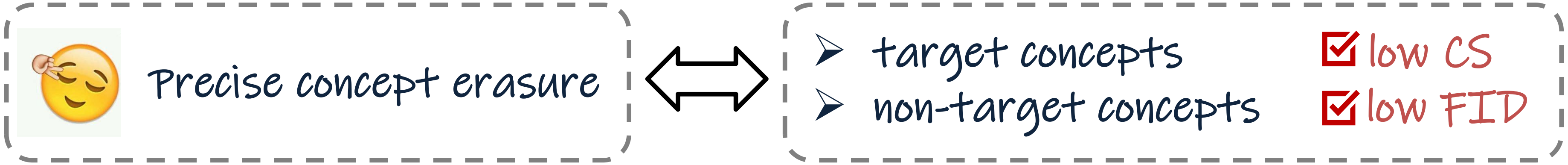
4. Experiments

■ Evaluation

- **Main Experiments**: AdaVD is applied to SD v1.4 to erase various concepts (specific instances, artistic styles, celebrities, NSFW concepts). Prompts are formatted with specific templates, generating 10 images per prompt under identical settings for quantitative and qualitative comparison.
- **Hyperparameter Analysis**: The effects of hyperparameters p, s, ε on erasure efficacy and prior preservation are analyzed using qualitative evaluations.
- **Transferability**: AdaVD is applied to SDXL v1.0 and other community versions to demonstrate its generalization.
- **Further Analysis**: Additional insights cover time efficiency, visualization of erased components, and downstream application potential.

■ Metrics

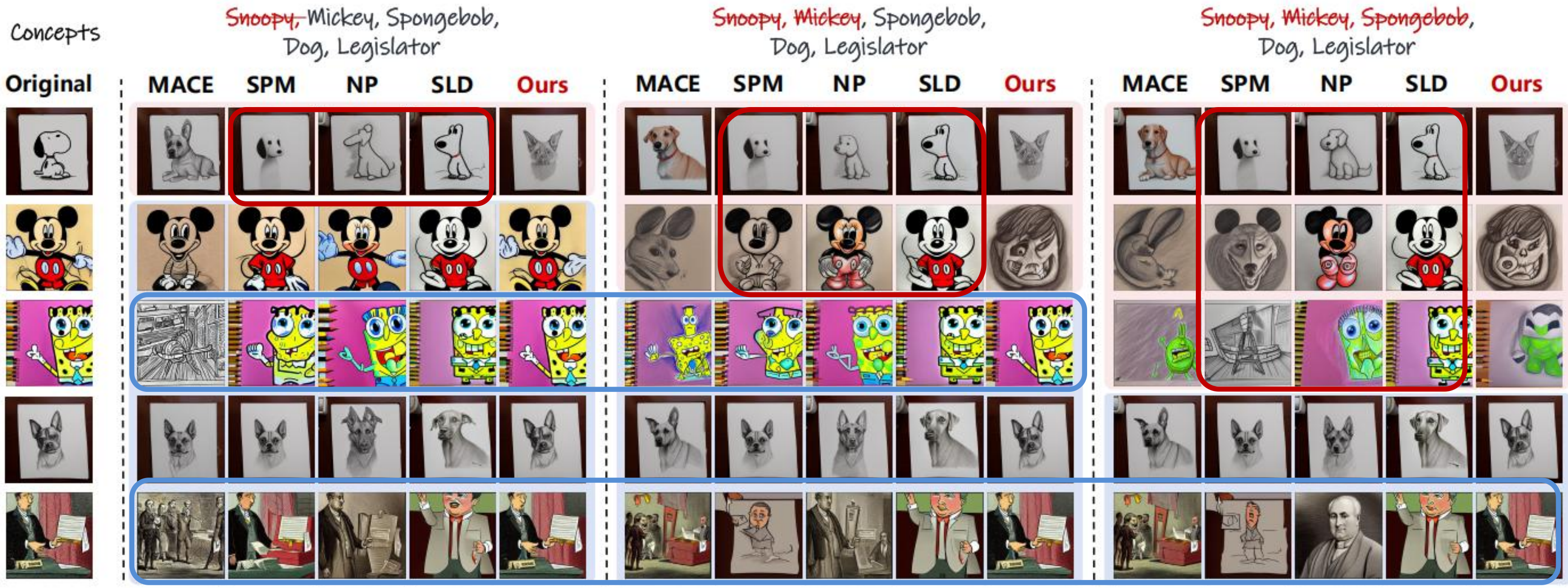
- **CLIP Score:** Measures erasure efficacy by calculating the similarity between the prompt containing the target concept and the generated images through CLIP encoder.
- **FID:** Evaluates prior preservation by quantifying the distributional distance between non-target concept images before and after erasure.



4. Experiments

■ Erasure Performance on SD v1.4

Specific Instance



Concept	Snoopy	Mickey	Spongebob	Pikachu	Dog	Legislator
	CS	CS	CS	CS	CS	CS
SD v1.4	28.51	26.57	27.43	-	-	-
			*Erase **Snoopy***			
	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓	FID ↓
ConAbl	25.38	38.44	41.59	29.68	27.76	27.36
MACE	20.78	118.01	111.90	81.99	43.27	65.97
SPM	23.89	33.06	34.70	23.89	19.61	18.26
NP	23.66	59.58	78.74	52.37	67.51	55.22
SLD	27.84	48.12	55.36	38.74	41.95	49.08
Ours	**20.28**	**5.72**	**8.56**	**5.79**	**2.32**	**6.07**
			*Erase **Snoopy** and **Mickey***			
	CS ↓	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓
ConAbl	24.26	24.08	46.32	39.63	30.57	27.49
MACE	20.74	20.71	51.49	110.67	52.07	77.13
SPM	23.16	22.81	41.58	31.77	21.96	23.69
NP	23.59	24.85	81.41	50.10	65.93	58.88
SLD	27.76	26.74	54.59	39.24	41.62	50.13
Ours	**20.29**	**19.93**	**9.34**	**5.84**	**2.41**	**6.43**
			*Erase **Snoopy** and **Mickey** and **Spongebob***			
	CS ↓	CS ↓	CS ↓	FID ↓	FID ↓	FID ↓
ConAbl	23.94	23.64	25.04	51.20	31.59	30.03
MACE	20.48	20.50	21.59	99.68	47.46	70.38
SPM	22.81	22.35	20.82	39.83	22.68	25.31
NP	24.29	24.76	25.31	64.75	65.10	59.33
SLD	27.84	26.71	27.60	39.41	42.32	49.88
Ours	**19.39**	**19.73**	**20.34**	**6.85**	**2.79**	**7.26**

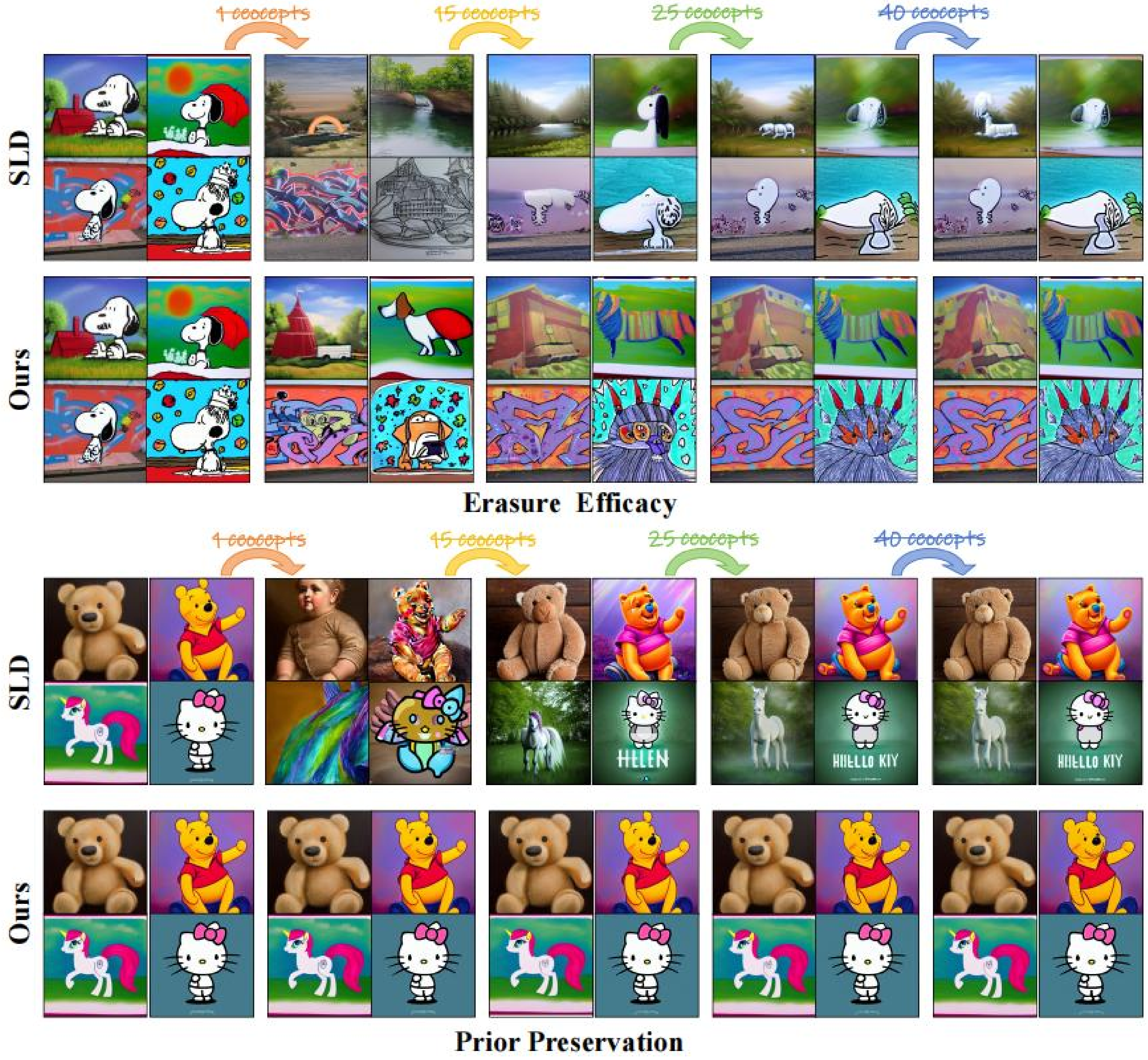
We evaluate AdaVD on both single- and multi-instance concept erasure tasks.

- AdaVD consistently achieves the lowest CS and FID across all cases, with its FID being less than 33% of the second-best method when erasing single instance concept.
- Moreover, AdaVD demonstrates superior performance in complex multi-concept erasure, maintaining the lowest CS and FID.

4. Experiments

■ **Erasure Performance on SD v1.4**

Multi Specific Instance

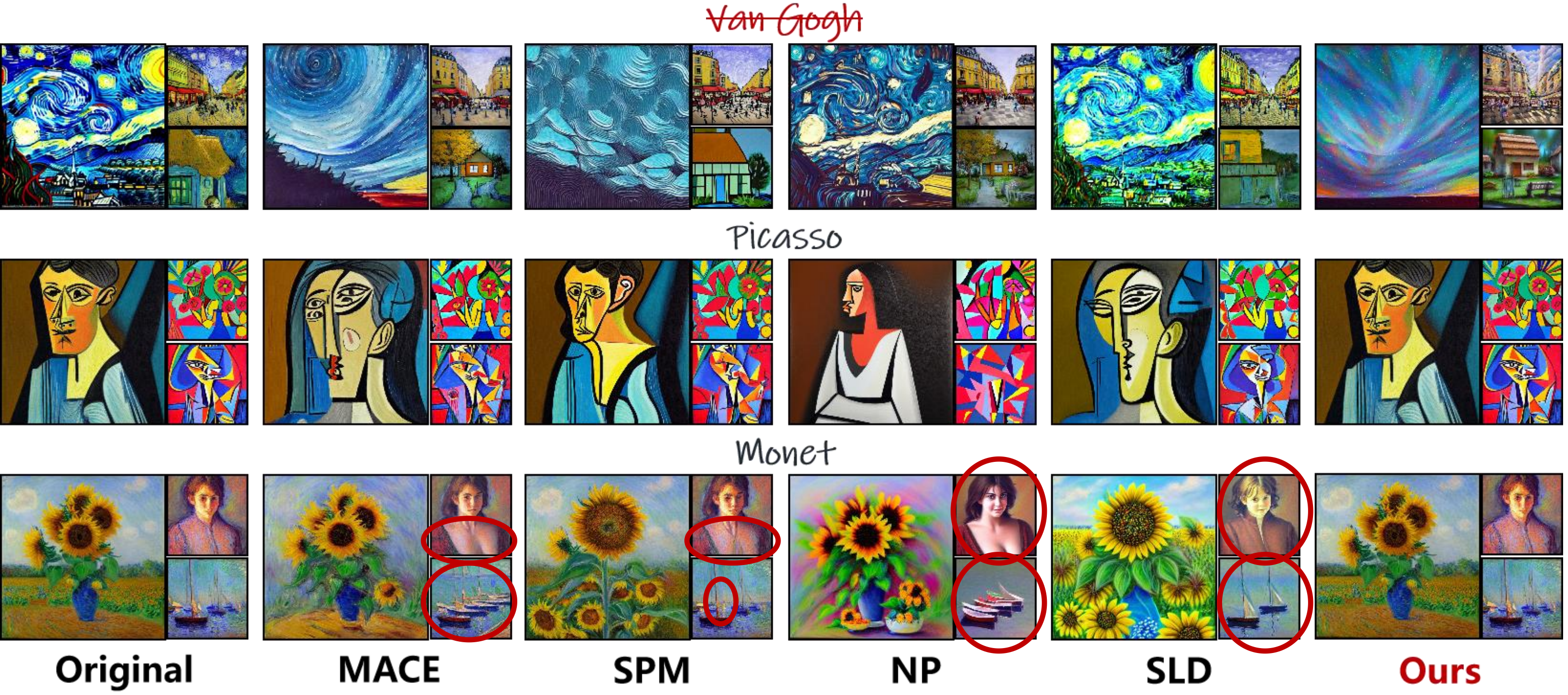


- SLD gradually loses its precision when erasing the target concepts.
  - ✗ SLD concatenates the target concepts making it hard to focus on each individual concept.
  - ✗ Additionally, some concepts may be truncated due to the token length limitation of the text encoder's tokenizer.
- AdaVD achieves consistently high performance in multi-concept erasure.
  - ✓ It constructs a value subspace based on the orthogonal complement of all the target concepts, which ensures that no information regarding any individual concept is lost.

4. Experiments

- **Erasure Performance on SD v1.4**

Art Style



We evaluate AdaVD on single-art style erasure task.

- Our AdaVD exhibits superior prior preservation, and achieves the lowest or closed-to-lowest CS and FID scores.
- Both MACE and SPM are effective in erasing the target concept, however, their prior preservation is somehow less satisfactory, which is particularly noticeable in the generated images in "Monet" style.

Concept	Van Gogh	Picasso	Monet	Andy Warhol	Caravaggio
	CS	CS	CS	CS	CS
SD v1.4	29.21	29.06	29.02	-	-
Erase Van Gogh					
	CS ↓	FID ↓	FID ↓	FID ↓	FID ↓
ConAbl	28.80	71.71	138.72	70.30	73.10
MACE	27.74	65.77	69.79	83.37	75.41
SPM	**24.78**	62.25	32.27	58.30	61.50
NP	24.90	141.56	124.52	127.85	136.32
SLD	27.48	103.96	109.11	103.89	119.32
Ours	24.87	**6.82**	**2.66**	**8.36**	**6.84**
Erase Picasso					
	FID ↓	CS ↓	FID ↓	FID ↓	FID ↓
ConAbl	58.62	27.72	140.34	73.35	67.44
MACE	60.46	27.11	49.92	76.10	72.85
SPM	38.79	26.69	7.76	52.00	51.40
NP	111.35	**26.14**	91.11	116.24	121.82
SLD	98.21	27.03	93.01	97.00	110.05
Ours	**5.49**	26.99	**2.33**	**9.38**	**7.05**
Erase Monet					
	FID ↓	FID ↓	CS ↓	FID ↓	FID ↓
ConAbl	141.52	132.10	24.53	208.38	186.26
MACE	76.90	69.35	26.89	88.35	81.72
SPM	41.03	29.71	27.00	31.90	25.99
NP	137.21	126.75	**24.47**	127.22	135.83
SLD	94.48	92.88	25.73	100.90	114.87
Ours	**6.94**	**6.50**	26.30	**8.46**	**7.19**

4. Experiments

■ Erasure Performance on SD v1.4

Celebrity

Concept	Bruce Lee	Marilyn Monroe	Melania Trump	Anne Hathaway	Tom Cruise
	CS	CS	CS	CS	CS
SD v1.4	30.77	27.67	29.80	-	-
Erase ***Bruce Lee***					
	CS	FID	FID	FID	FID
ConAbl	31.35	57.79	40.95	48.08	53.53
MACE	25.04	74.80	68.83	75.05	71.20
SPM	27.75	26.89	7.83	9.46	28.54
NP	24.70	102.67	82.13	89.60	89.92
SLD	28.22	87.15	84.32	85.37	94.07
Ours	**20.67**	**6.68**	**5.08**	**6.39**	**13.11**
Erase ***Marilyn Monroe***					
	FID	CS	FID	FID	FID
ConAbl	66.97	28.75	51.52	58.57	54.13
MACE	76.23	**19.52**	71.05	74.90	73.06
SPM	32.70	21.87	25.27	22.86	19.34
NP	113.12	25.86	87.27	98.86	86.70
SLD	87.83	26.70	107.42	102.13	81.12
Ours	**7.88**	19.87	**4.46**	**5.43**	**9.33**
Erase ***Melania Trump***					
	FID	FID	CS	FID	FID
ConAbl	54.46	59.10	29.89	58.65	54.50
MACE	78.07	71.34	**20.71**	73.49	71.09
SPM	14.08	30.40	23.12	28.85	22.35
NP	115.35	103.83	23.73	106.04	106.00
SLD	90.69	93.93	25.45	104.48	88.31
Ours	**7.32**	**6.86**	23.28	**6.52**	**5.74**

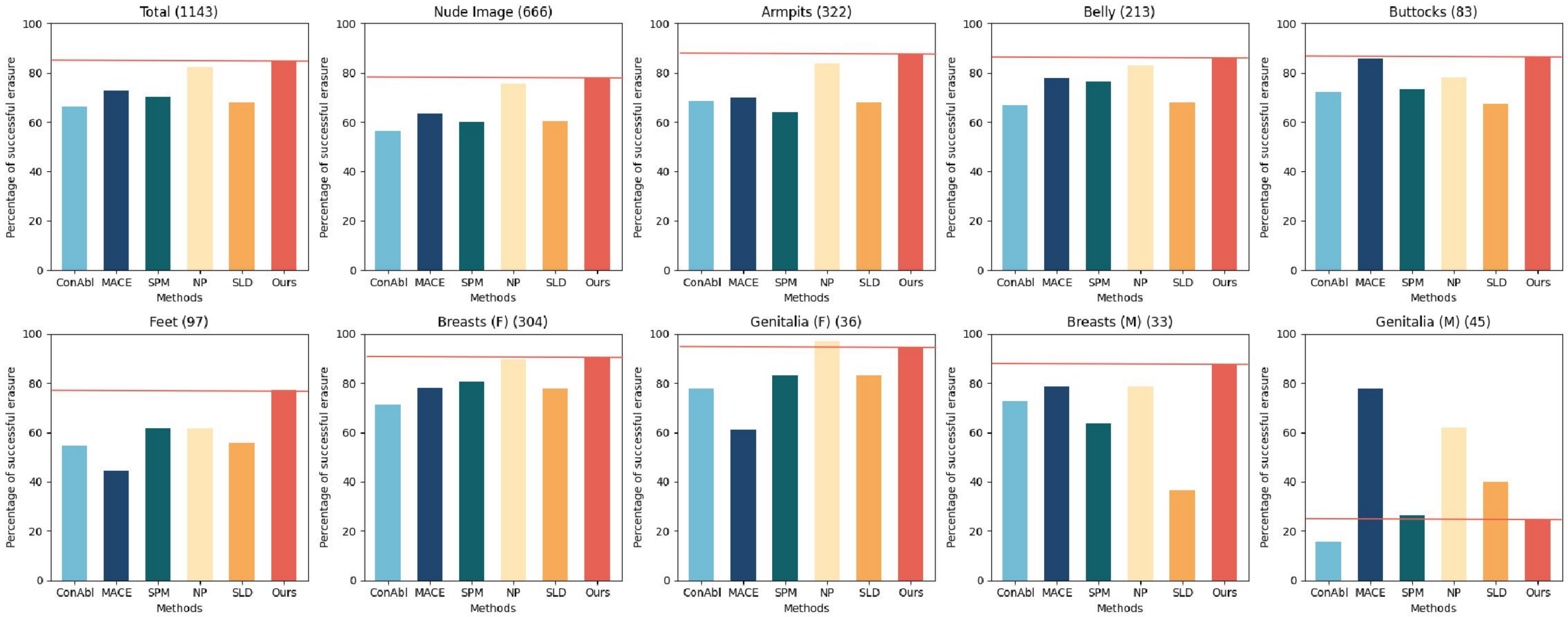
We evaluate AdaVD on single-celebrity erasure task.

- AdaVD achieves the lowest or near-lowest CS and FID values, parti-cularly excelling in FID.
- For non-target concepts, all the four competing methods have caused some quite strong deviations, altering the original images. This is particularly noticeable in the generated images from the prompt corre-sponding to “Melania Trump”.

4. Experiments

- **Erasure Performance on SD v1.4**

NSFW Concept



We experiment with erasing the "nudity" concept using the I2P benchmark, and detecting nude items by NudeNet. AdaVD outperforms both training-based and training-free methods, achieving a near 85% removal rate and the highest success rate in most categories.

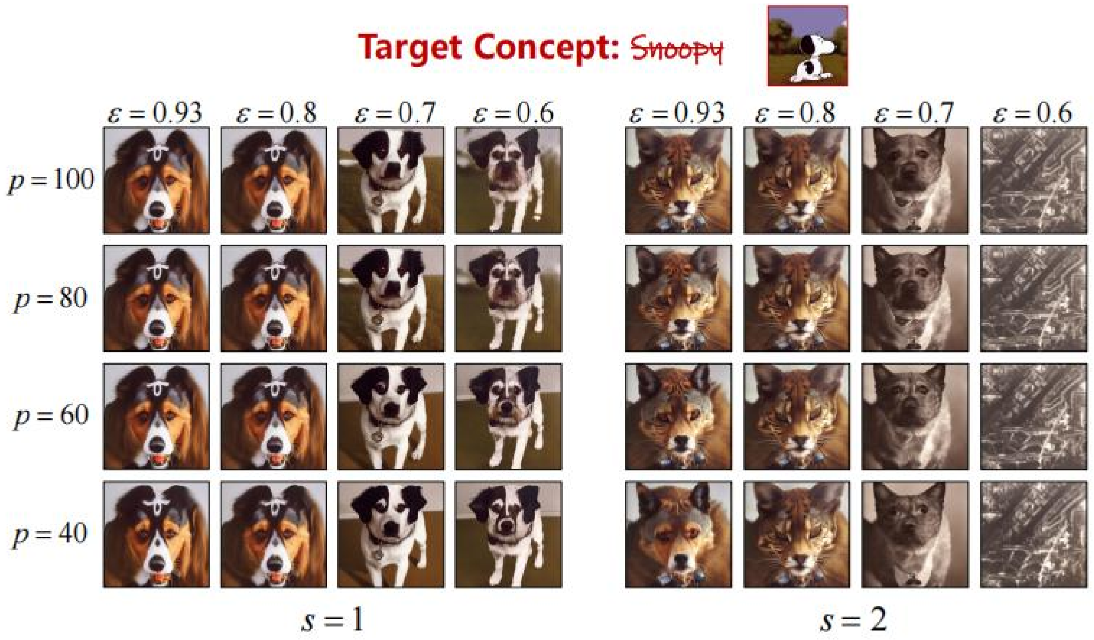
4. Experiments

$$\delta(\boldsymbol{x}, \boldsymbol{y}) = \frac{s}{1 + e^{-p(\cos(\boldsymbol{x}, \boldsymbol{y}) - \epsilon)}}.$$

- ε: quantify and filter
- s: control the factor scale
- p: control the increasing rate

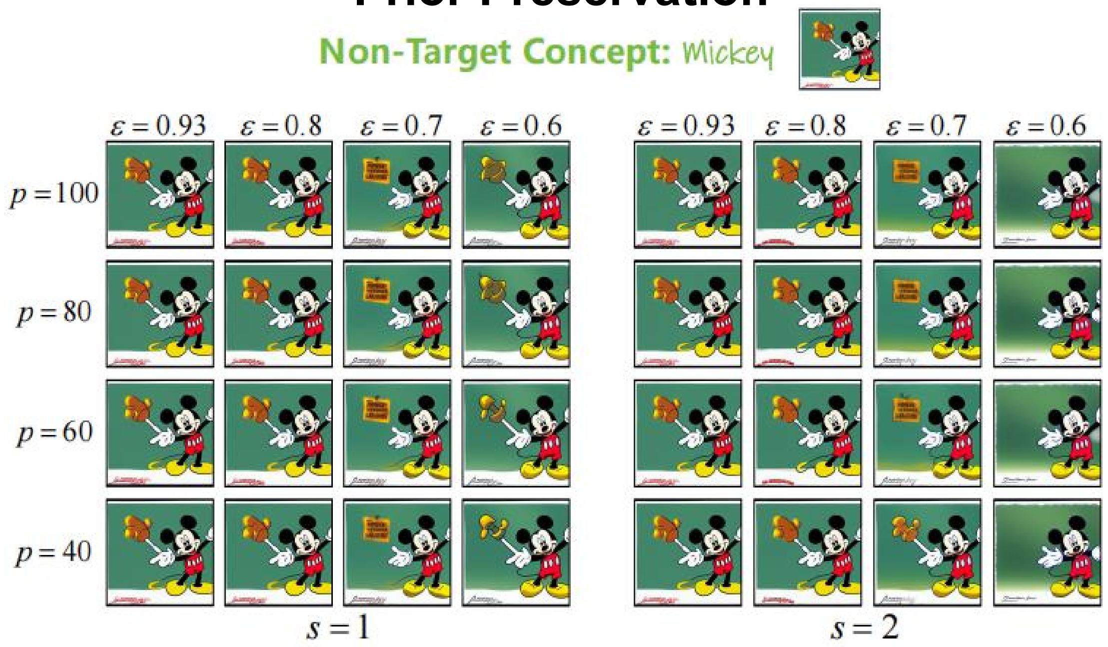
■ Hyper-parameter Analysis

Erasure Efficacy



- ✓ **Impact of ε**: A lower threshold ε enhances erasure efficacy.
- ✓ **Impact of s**: A larger factor scale s amplifies erasure efficacy, but excessive erasure occurs when s = 2 and ε = 0.6
- ✓ **Impact of p**: p has a milder influence on erasure efficacy.

Prior Preservation

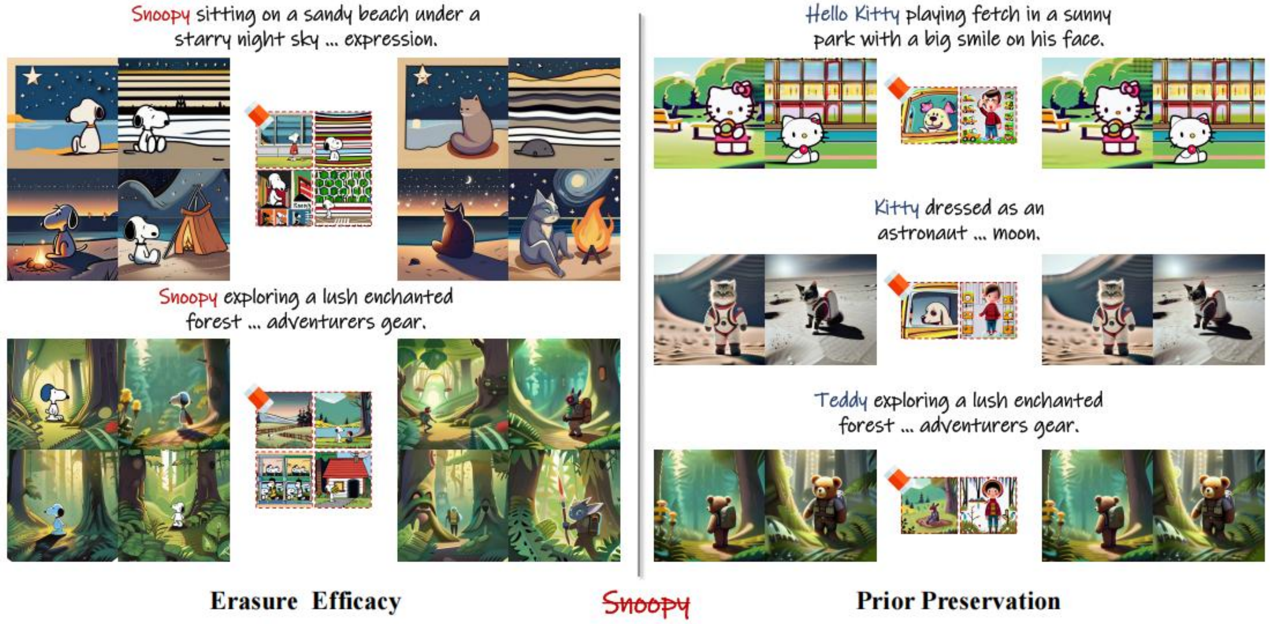


- ✓ **Impact of ε**: A lower ε harms non-target concept generation.
- ✓ **Impact of s**: A larger s increases deviations in non-target concept generation due to amplified token shifts.
- ✓ **Impact of p**: A lower p reduces deviations in non-target concept generation at higher s. Conversely, at s=1, a higher p enhances the preservation of related non-target concepts.

4. Experiments

■ Transferability to Other T2I Models

On Community Versions



On SDXL v1.0

■ Time Consumption

	Data Preparation	Model Finetune	Image Generation	Total Time
ConAbl	9290	1120	0.9	10419
SPM	0	72850	1.7	72867
MACE	303	232	0.9	544
SLD	0	0	1.4	14
Ours	4	0	1.8	22

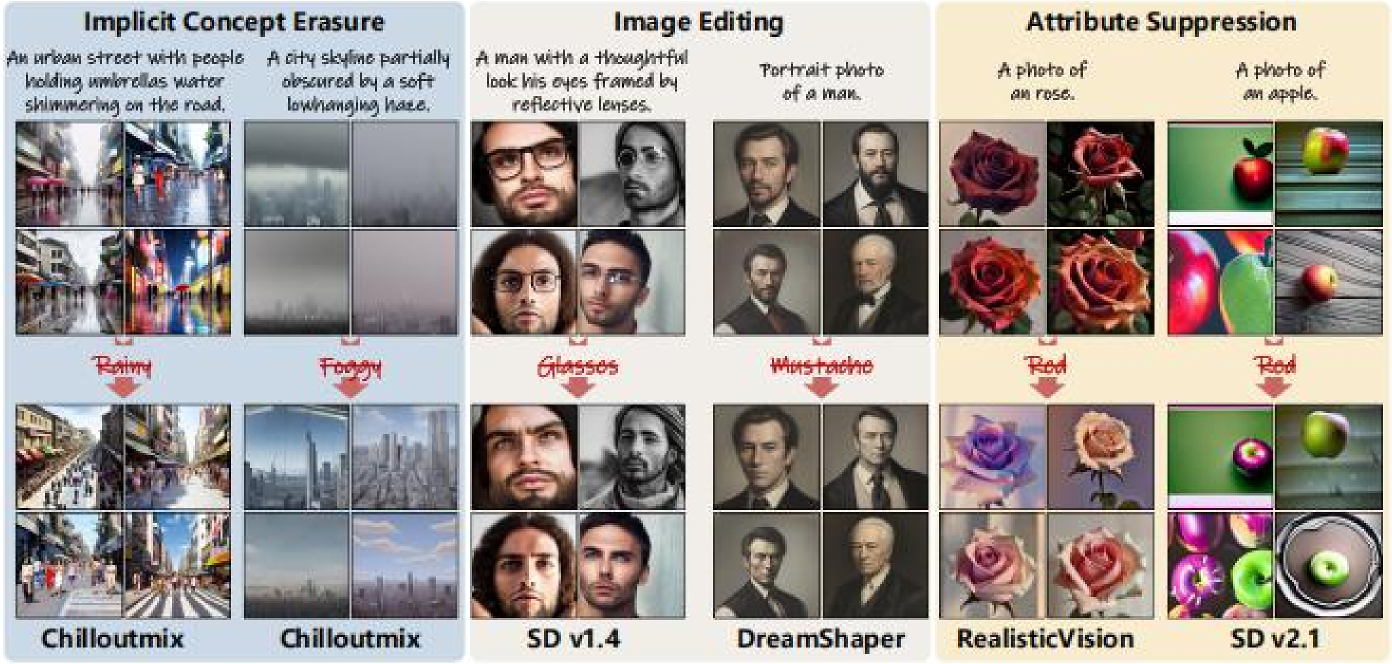
- The two training-free methods of SLD and AdaVD are significantly faster as no fine-tuning is needed.
- Our AdaVD costs slightly more time than SLD, i.e., a total of 8 extra seconds due to its basis computation. But this mild increase yields a significant performance gain, succeeding in precise concept erasure.

4. Experiments

Interpreting Erased Components by Visualization

- The erased components consistently align with the corresponding target semantics when dealing with the target concepts.
- The erased components do not contain any informative pattern for non-target concepts, indicating that they carry no meaningful semantics.

Downstream Applications:



- **Implicit Concept Erasure**: Effectively handles implicit concept erasure, as evidenced by the removal of "rainy" and "foggy" without explicit mentions.
- **Image Editing**: Precisely erases appearance concepts such as "glasses" and "mustache" with minimal impact on other details.
- **Attribute Suppression**: Eliminates strongly coupled attributes, e.g., the color "red" from objects like apples and roses.

Thanks for your listening!

Presented by: Yuan Wang